

4장. 표집방법(Book2)

1절. 표집(표본추출 sampling)

표본조사: 모집단의 특성을 조사하기 위하여 모집단의 일부(표본)를 추출하여 모집단의 특성을 추론↔전수조사,

표본추출(표집, sampling): 표본을 추출하는 방법 및 과정

2절. 표집의 기본 용어

1) 모집단: 특징을 조사하기 위한 전체 대상

2) 표집단위(sampling unit): 표본으로 추출되는 모집단의 구성분자(조사단위: 정보수집의 기본)

(관찰단위: 편의적으로 표본으로 추출되는 모집단의 구성분자)

4) 표집간격(sampling interval), 표집률(sampling ratio)

1개의 표집단위가 추출되는 모집단의 구성분자의 수 & 모집단에서 구성분자가 표본으로 선택될 비율=1/표집간격

eg) 모집단크기=1000, 표본크기=50 → 표집간격=1000/50=20 & 표집률=1/20=0.05

6) 모수(parameter): 수치적으로 표현가능한 특성

7) 통계량: 표본을 이용해서 만든 함수

8) 표집틀(frame): 추출단위의 목록

3절. 표집(표본추출)의 종류 - 확률표집 & 비확률표집

1) 확률표집

각 표집단위가 표본에 추출될 확률이 동일, 신뢰구간등 특성에 대한 통계적 추론이 가능.

①단순무작위(임의)표집(simple random sampling)

크기 N 인 모집단으로부터 크기 n 인 표본을 뽑을 경우, 모든 가능한 크기 n 인 표본이 동일하게 선출될 기회를 갖는 추출방법

개념상 간단하다.

표본원소들이 광범위한 지역에 흩어지게 된다.

조사비용이 많이 상승한다.

②계통추출법(systematic sampling)

표집틀(frame)(구해진 경우)에서 구성요소에 대해 일정한 순서대로 추출

eg)110명에서 10명을 추출하는 경우: $110/10=11$, (1~11)에서 3번이 임의로 추출되었다면

표본=(3, $3+11$, $3+11\times 2$, ... , $3+11\times 9$)=(3,14,25,..., 102)

③층화추출법(stratified random sampling)

모집단을 동질적인 몇 개의 층으로 나눈 후 각 층에서 임의추출(무작위표집)

각층내 등질적, 층간 이질적.

④집락추출법(cluster sampling)

모집단을 여러개의 집락으로 나누어 집락을 1차추출단위로 하여 단순임의추출후 추출된 집락에서 단순임의추출.

각집락내 이질적, 집락간 등질적.

2) 비확률표집

표본크기가 작은 경우 유리, 시간과 비용이 적게 듦, 개략적인 정보가 필요한 경우.

①편의(convenience)표집

정해진 표본크기가 될 때까지 표집

eg) 앙케이트조사(같은 질문을 여러 사람에게 제시하여 의견구함, TV)

②유의(purposive)표집(판단(judgement))

전문가의 판단으로 표본추출

③할당(quota)표집

모집단 특성(인구통계적 특성) 비율에 맞추어 소표본의 크기를 할당

④눈덩이(snowball)표집

소수의 인원을 표본으로 추출하여 조사 후, 그 소수의 인원이 표본을 추출 -반복

모집단구성원을 찾기 어려운 경우.

4절. 표집오차와 표본의 크기

표집오차(표본추출오차,sampling error):표집중에 발생하는 오차(전수조사가 아니어서 발생하는 오차, $|\hat{\mu} - \mu|$)

표준오차((standard error)SRS)= $\sqrt{s^2/n}$

비표집오차(비표본추출오차): 표집중이외에서 발생하는 오차~ 측정상 오차

단순임의표집에서 $(1 - \alpha)100\%$ 신뢰수준에서 최대허용오차($e = \max|\bar{y} - \mu|$)가 주어진 경우

표본의 크기(SRS)= $\left\lceil \left(\frac{z_{\alpha} \sigma}{e} \right)^2 \right\rceil$,

$z_{0.9} = 1.645$, $z_{0.95} = 1.96$, $\lceil \rceil$ =ceil함수,

σ =모표준편차 or 과거 조사에서 구해진 모표준편차의 추정값

eg) A공단 근로자의 임금추정 표본조사. 95% 신뢰수준에서 추정오차가 10만원 이내로 하기 위해서 필요한 표본의 크기? ($\sigma = 50$ 만원)

$$n = \left\lceil \left(\frac{z_{\alpha} \sigma}{e} \right)^2 \right\rceil = \left\lceil \left(\frac{1.96 \times 50}{10} \right)^2 \right\rceil = [96.04] = 97 \text{명}$$

신뢰구간:

*신뢰도가 클수록 신뢰구간이 더 넓다.

*표본크기가 클수록 신뢰구간이 더 좁다.

4.1 단순임의 추출법(Simple Random Sampling)

크기 N 인 모집단으로부터 크기 n 인 표본을 뽑을 경우, 모든 가능한 크기 n 인 표본이 동일하게 선출될 확률 $\left[\frac{1}{\binom{N}{n}} \right]$ 를 갖는 추출법(비복원추출).

1. 단순임의추출법의 특징

① 개념상 간단하다. ① 표본원소들이 광범위한 지역에 흩어지게 된다. ② 조사비용이 많이 상승한다.

2. 모평균(μ), 표본평균($\bar{y}$), 모총합(τ)

모집단: $\{Y_1, \dots, Y_N\} \Rightarrow$ 모평균: $\mu = \frac{1}{N} \sum_{i=1}^N Y_i$ ($\tau = \sum_{i=1}^N Y_i$: 모총합)

표본: $\{y_1, \dots, y_n\} \Rightarrow$ 표본평균: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, 모총합의 추정값: $\hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y}$

$\hat{\mu} = \bar{y}$ & $\bar{y}$ 는 μ 의 불편추정량(unbiased estimator)이다. ($E(\bar{y}) = \mu$)

모집단= $\{Y_1, \dots, Y_N\}$, 표본= $\{y_1, y_2, \dots, y_n\}$, $y_i = Y_j$ for $i \in (1, \dots, n)$ and $j \in (1, \dots, N)$

3. 모분산(S^2, σ^2)

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \mu)^2, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2$$

4. $\bar{y}$ 의 분산($V(\bar{y})$)

$$V(\bar{y}) = (1 - \frac{n}{N}) \frac{S^2}{n} \text{ (비복원 추출)}, \quad V(\bar{y}) = \frac{\sigma^2}{n} \text{ (복원 추출)},$$

$$\frac{N-n}{N} = 1 - \frac{n}{N} : \text{분산의 유한모집단수정(finite population correction: fpc)}$$

$$\frac{n}{N} : \text{추출률(sampling fraction)}$$

$$\text{만일 } \frac{n}{N} \leq 0.05 \text{ 이면, fpc 무시 가능 } \sim \frac{N-n}{N-1} = \frac{N-n}{N} \approx 1$$

5. $V(\bar{y})$ 의 추정($\hat{V}(\bar{y})$)

$$\text{표본분산: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{V}(\bar{y}) = (1 - \frac{n}{N}) \frac{s^2}{n} \text{ (비복원 추출)}, \quad \hat{V}(\bar{y}) = \frac{s^2}{n} \text{ (복원 추출)}$$

$$6. \mu \text{에 대한 신뢰구간 : } \bar{y} \pm z_{\alpha/2} \sqrt{\hat{V}(\bar{y})}$$

eg) 모집단= $\{Y_1, \dots, Y_5\} = \{1, 2, 3, 4, 5\}$, $N=5$, 표본= $\{y_1, y_2, y_3\}$, $n=3$

가능한 표본의 수= $\binom{N}{n} = \binom{5}{3} = 10$ 개 ~ $\{1, 2, 3\}, \{1, 2, 4\}, \dots, \{3, 4, 5\}$ ~ 각 표본이 실제로 뽑힐 확률=1/10

$$\tau = 1 + 2 + 3 + 4 + 5 = 15, \quad \mu = \tau / N = 15 / 5 = 3$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2 = \frac{1}{5} [(1-3)^2 + \dots + (5-3)^2] = \frac{10}{5} = 2$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \mu)^2 = \frac{1}{5-1} [(1-3)^2 + \dots + (5-3)^2] = \frac{10}{4} = 2.5$$

$$V(\bar{y}) = (1 - \frac{n}{N}) \frac{S^2}{n} = (1 - \frac{3}{5}) \frac{2.5}{3} = 0.333 (\text{비복원 추출}),$$

if 표본= $\{y_1, y_2, y_3\}=\{1, 3, 4\}$, $\bar{y} = (1+3+4)/3 = 2.667$,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{2} [(1-2.667)^2 + (3-2.667)^2 + (4-2.667)^2] = \frac{1}{2} \times 4.667 = 2.333$$

$$\hat{V}(\bar{y}) = (1 - \frac{n}{N}) \frac{s^2}{n} = (1 - \frac{3}{5}) \frac{2.333}{3} = 0.311 (\text{if 비복원 추출한 표본이 } \{y_1, y_2, y_3\}=\{1, 3, 4\}),$$

μ 에 대한 95% 신뢰구간 (if 비복원 추출한 표본이 $\{y_1, y_2, y_3\}=\{1, 3, 4\}$):

$$(\bar{y} \pm z_{\alpha/2} \sqrt{\hat{V}(\bar{y})}) = (2.667 \pm 1.96 \sqrt{0.311}) = (2.667 \pm 0.399) = (2.268, 3.066)$$

eg) 40명 신입사원의 평균 월급여(μ) & 전체 월급여(τ) 조사

4명을 임의 추출(비복원) $\rightarrow N=40, n=4$, 표본= $(y_1, y_2, y_3, y_4)=(230, 260, 240, 250)$

y_i : 추출된 4명중 i 번째 사람의 월급여, $i=1, \dots, 4$

$$\bar{y} = \frac{1}{4} (230 + 260 + 240 + 250) = 245 \text{ 만}, \quad \hat{\tau} = N\bar{y} = 40 \times 245 = 9800 \text{ 만}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{3} [(-15)^2 + 15^2 + (-5)^2 + 5^2] = 125$$

$$\hat{V}(\bar{y}) = (1 - \frac{n}{N}) \frac{s^2}{n} = (1 - \frac{4}{40}) \frac{125}{4} = 28.125 (\text{since 비복원 추출}),$$

μ 에 대한 95% 신뢰구간:

$$(\bar{y} \pm z_{\alpha/2} \sqrt{\hat{V}(\bar{y})}) = (245 \pm 1.96 \sqrt{28.125}) = (245 \pm 5.303) = (239.697, 250.303)$$

층화임의 추출법(Stratified Random sampling)

모집단을 어떤 층화기준에 의하여 여러 개의 층으로 분할 한 다음 각층에서 독립적으로 일정한 수의 표본을 임의추출하는 방법. 모집단에 속하는 모든 단위를 어떤 지표에 의하여 몇 개의 집단으로 분할하되 분할된 집단의 각 단위들이 등질적인 것이 되도록 하는 것을 층화(層化, stratification), 그 분할된 집단을 층(層, stratum). 층화할 때는 층내는 등질적(等質的, homogeneous)이 되도록 하고 층간에는 이질적(異質的, heterogeneous)이 되도록 하여야 함 - 한 층에 속하는 단위들은 같은 성질인 것끼리 모아야 하고 층과 층은 서로 다른 성질이 되도록.

집락표본 추출방법(Cluster Random Sampling)

모집단을 기본단위들로 묶은 집락(cluster)으로 구성하고, 모집단에 집락을 1차추출단위로 임의추출하고, 추출된 각 집락에서 표본을 임의추출하는 방법. 집락내에서는 이질적이 되어야 하고 하나의 집락은 집락 추출의 경제적 장점을 살리기 위해 다른 집락과 등질적이어야 함. 집락=PSU(1차추출단위), 실제조사단위=SSU(2차추출단위)